Annotating Opinions in the World Press

Theresa Wilson
Intelligent Systems Program
University of Pittsburgh
Pittsburgh, PA 15260, USA
twilson@cs.pitt.edu

Janyce Wiebe
Department of Computer Science
University of Pittsburgh
Pittsburgh, PA 15260, USA
wiebe@cs.pitt.edu

Abstract

In this paper we present a detailed scheme for annotating expressions of opinions, beliefs, emotions, sentiment and speculation (private states) in the news and other discourse. We explore inter-annotator agreement for individual private state expressions, and show that these low-level annotations are useful for producing higher-level subjective sentence annotations.

1 Introduction

In this paper we present a detailed scheme for annotating expressions of opinions, beliefs, emotions, sentiment, speculation and other private states in newspaper articles. *Private state* is a general term that covers mental and emotional states, which cannot be directly observed or verified (Quirk et al., 1985). For example, we can observe evidence of someone else being happy, but we cannot directly observe their happiness. In natural language, opinions, emotions and other private states are expressed using subjective language (Banfield, 1982; Wiebe, 1994).

Articles in the news are composed of a mixture of factual and subjective material. Writers of editorials frequently include facts to support their arguments, and news reports often mix segments presenting objective facts with segments presenting opinions and verbal reactions (van Dijk, 1988). However, natural language processing applications that retrieve or extract information from or that summarize or answer questions about news and other discourse have focused primarily on factual information and thus could benefit from knowledge of subjective language. Traditional information extraction and information retrieval systems could learn to concentrate on objectively presented factual information. Question answering systems could identify when an answer is speculative rather than certain. In addition, knowledge of how opinions and other private states are realized in text would directly support new tasks, such as opinion-oriented information extraction (Cardie et al., 2003). The ability to extract opinions when they appear in documents would benefit multi-document summarization systems seeking to summarize different opinions and perspectives, as well as multi-perspective question-answering systems trying to answer opinion-based questions.

The annotation scheme we present in this paper was developed as part of a U.S. government-sponsored project (ARDA AQUAINT NRRC)[1] to investigate multiple perspectives in question answering (Wiebe et al., 2003). We implemented the scheme in GATE[2], a General Architecture for Text Engineering (Cunningham et al., 2002). General instructions for annotating opinions and specific instructions for downloading and using GATE to perform the annotations are available at

[1]This work was performed in support of the Northeast Regional Research Center (NRRC) which is sponsored by the Advanced Research and Development Activity in Information Technology (ARDA), a U.S. Government entity which sponsors and promotes research of import to the Intelligence Community which includes but is not limited to the CIA, DIA, NSA, NIMA, and NRO.

[2]GATE is freely available from the University of Sheffield at http://gate.ac.uk.

http://www.cs.pitt.edu/˜wiebe/pubs/ardasummer02. The annotated data will be available to U.S. government contractors this summer. We are working to resolve copyright issues to make it available to the wider research community.

In developing this annotation scheme, we had two goals. The first was to develop a representation for opinions and other private states that was built on work in linguistics and literary theory on subjectivity (please see (Banfield, 1982; Fludernik, 1993; Wiebe, 1994; Stein and Wright, 1995) for references). The study of subjectivity in language focuses on how private states are expressed linguistically in context. Our second goal was to develop an annotation scheme that would be useful for corpus-based research on subjective language and for the development of applications such as multi-perspective question-answering systems. The annotation scheme that resulted is more detailed and comprehensive than previous ones for subjective language.

Our study of the annotations produced by the annotation scheme gives two important results. First, we find that trained annotators can consistently perform detailed opinion annotations with good agreement (0.81 Kappa). Second, the agreement results are better than in previous sentence-level annotation studies, suggesting that adding detail can help the annotators perform more reliably.

In the sections that follow, we first review how opinions and other private states are expressed in language (section 2) and give a brief overview of previous work in subjectivity tagging (section 3). We then describe our annotation scheme for private state expressions (section 4) and give the results of an annotation study (section 5). We conclude with a discussion of our findings from the annotation study and future work (section 6). In the appendix, we give sample annotations as well as a snapshot of the annotations in GATE.

2 Expressing Private States in Text

2.1 Private States, Speech Events, and Expressive Subjective Elements

There are two main ways that private states are expressed in language. Private states may be explicitly mentioned, or they may be expressed indirectly by the types of words and the style of language that a speaker or writer uses. An example of an explicitly-mentioned private state is "frustrated" in sentence (1).

> **(1)** Western countries were left frustrated and impotent after Robert Mugabe formally declared that he had overwhelmingly won Zimbabwe's presidential election.

Although most often verbs, it is interesting to note that explicit mentions of private states may also be nouns, such as "concern" in "international concern" and "will" in "will of the people." They may even be adjectives, such as "fearful" in "fearful populace."

The second way that private states are generally expressed is indirectly using *expressive subjective elements* (Banfield, 1982). For example, the private states in sentences (2) and (3) are expressed entirely by the words and the style of language that is used.

> **(2)** The time has come, gentlemen, for Sharon, the assassin, to realize that injustice cannot last long.

> **(3)** "We foresaw electoral fraud but not daylight robbery," Tsvangirai said.

In (2), although the writer does not explicitly say that he hates Sharon, his choice of words clearly demonstrates a negative attitude. In sentence (3), describing the election as "daylight robbery" clearly reflects the anger being experienced by the speaker, Tsvangirai. As used in these sentences, the phrases "The time has come," "gentlemen," "the assassin," "injustice cannot last long," "fraud," and "daylight robbery" are all expressive subjective elements. Expressive subjective elements are used by people to express their frustration, anger, wonder, positive sentiment, mirth, etc., without explicitly stating that they are frustrated, angry, etc. Sarcasm and irony often involve expressive subjective elements.

When looking for opinions and other private states in text, an annotator must consider speech events as well as explicitly-mentioned private states. In this work, we use *speech event* to refer to any event of speaking or writing. However, the mere presence of a speech event does not indicate a private state. Both sentences (3) above and (4) below contain speech events indicated by "said." As mentioned previously, sentence (3) is opinionated, while in (4) the information is presented as factual.

> **(4)** Medical Department head Dr Hamid Saeed said the patient's blood had been sent to the Institute for Virology in Johannesburg for analysis.

For speech terms such as "said," "added," "told," "announce," and "report," an annotator determines if there is a private state mainly by looking inside the scope of the speech term for expressive subjective elements.

Occasionally, we also find private states that are expressed by direct physical actions. We call such actions *private state actions*. Examples are booing someone, sighing heavily, shaking ones fist angrily, waving ones hand dismissively, and frowning. "Applauding" in sentence (5) is an example of a positive-evaluative private state action.

> **(5)** As the long line of would-be voters marched in, those near the front of the queue began to spontaneously applaud those who were far behind them.

2.2 Nested Sources

An important aspect of a private state or speech event is its source. The source of a speech event is the speaker or writer. The source of a private state is the experiencer of the private state, i.e., the person whose opinion or emotion is being expressed. Obviously, the writer of an article is a source, because he wrote the sentences composing the article, but the writer may also write about other people's private states and speech events, leading to multiple sources in a single sentence. For example, each of the following sentences has two sources: the writer (because he wrote the sentences), and Sue (because she is the source of a speech event in (6) and of private states in (7) and (8), namely thinking and being afraid).

> **(6)** Sue said, "The election was fair."
> **(7)** Sue thinks that the election was fair.
> **(8)** Sue is afraid to go outside.

Note, however, that we don't really know what Sue says, thinks or feels. All we know is what the writer tells us. Sentence (6), for example, does not directly present Sue's speech event but rather Sue's speech event according to the writer. Thus, we have a natural *nesting of sources* in a sentence.

The nesting of sources may be quite deep and complex. For example, consider sentence (9).

> **(9)** The Foreign Ministry said Thursday that it was "surprised, to put it mildly" by the U.S. State Department's criticism of Russia's human rights record and objected in particular to the "odious" section on Chechnya.

There are three sources in this sentence: the writer, the Foreign Ministry, and the U.S. State Department. The writer is the source of the overall sentence. The remaining explicitly mentioned private states and speech events in (9) have the following nested sources:

said: (writer, Foreign Ministry)
surprised, to put it mildly: (writer, Foreign Ministry, Foreign Ministry)
criticism: (writer, Foreign Ministry, Foreign Ministry, U.S. State Dept.)
objected: (writer, Foreign Ministry)

Expressive subjective elements may also have nested sources. In sentence (9), "to put it mildly" and "odious" are expressive subjective elements, both with nested source (writer, Foreign Ministry). We might expect that an expressive subjective element always has the same nested source as the immediately dominating private state or speech term. Although this is the case for "odious" in (9) (the nested source of "odious" and "objected" is the same), it is not the same for "bigger than Jesus" in (10):

> **(10)** "It is heresy," said Cao. "The 'Shouters' claim they are bigger than Jesus."

The nested source of the subjectivity expressed by "bigger than Jesus" is Cao, while the nested source of "claim" is (writer, Cao, Shouters).[3]

[3](10) is an example of a *de re* rather than *de dicto* propositional attitude report (Rapaport, 1986).

3 Previous Work on Subjectivity Tagging

In previous work (Wiebe et al., 1999), a corpus of sentences from the Wall Street Journal Treebank Corpus (Marcus et al., 1993) was manually annotated with subjectivity classifications by multiple judges. The judges were instructed to classify a sentence as subjective if it contained any significant expressions of subjectivity, attributed to either the writer or someone mentioned in the text, and to classify the sentence as objective, otherwise. The judges rated the certainty of their answers on a scale from 0 to 3.

Agreement in the study was summarized in terms of Cohen's Kappa (κ) (Cohen, 1960), which compares the total probability of agreement to that expected if the taggers' classifications were statistically independent (i.e., "chance agreement"). After two rounds of tagging by three judges, an average pairwise κ value of 0.69 was achieved on a test set. On average, the judges rated 15% of the sentences as very uncertain (rating 0). When these sentences are removed, the average pairwise κ value is 0.79. When sentences with uncertainty judgment 0 or 1 are removed (on average 30% of the sentences), the average pairwise κ is 0.88.

4 An Annotation Scheme for Private States

The annotation scheme described in this section is more detailed and comprehensive the previous ones for subjective language. In (Wiebe et al., 1999), summary subjective/objective judgments were performed at the sentence level. For this work, annotators are asked to mark within each sentence the word spans that indicate speech events or that are expressions of private states. For every span that an annotator marks, there are a number of attributes the annotator may set to characterize the annotation.

The annotation scheme has two main components. The first is an annotation type for explicitly-mentioned private states and speech events. The second is an annotation type for expressive subjective elements. Table 1 lists the attributes that may be assigned to these two types of annotations. In addition, there is an annotation type, *agent*, that annotators may use to mark the noun phrase (if one exists) of the source of a private state or speech event.

Explicit private states/speech events
nested-source
onlyfactive: *yes, no*
overall-strength: *low, medium, high, extreme*
on-strength: *neutral, low, medium, high, extreme*
attitude-type: *positive, negative, both* (exploratory)
attitude-toward (exploratory)
is-implicit
minor
Expressive subjective elements
nested-source
strength: *low, medium, high, extreme*
attitude-type: *positive, negative, other* (exploratory)

Table 1: Attributes for the two main annotation types. For attributes that take on one of a fixed set of values, the set of possible values are given.

4.1 Explicitly-mentioned Private State and Speech Event Annotations

An important part of the annotation scheme is represented by the *onlyfactive* attribute. This attribute is marked on every private state and speech event annotation. The *onlyfactive* attribute is used to indicate whether the source of the private state or speech event is indeed expressing an emotion, opinion or other private state. By definition, any expression that is an explicit private state (e.g., "think", "believe," "hope," "want") or a private state mixed with speech (e.g., "berate," "object," "praise") is *onlyfactive=no*. On the other hand, *neutral* speech events (e.g., "said," "added," "told") may be either *onlyfactive=yes* or *onlyfactive=no*, depending on their contents. For example, the annotation for "said" in sentence (3) would be marked *onlyfactive=no*, but the annotation for "said" in sentence (4) would be marked *onlyfactive=yes* (sentences in section 2).

Note that even if *onlyfactive=no*, the sentence may express something the nested source believes is factual. Consider the sentence "John criticized Mary for smoking." John expresses a private state (his negative evaluation of Mary's smoking). However, this does not mean that John does not believe that Mary smokes.

Like the *onlyfactive* attribute, the *nested-source* attribute is included on every private state and speech event annotation. The nested source (i.e., (writer, Foreign Ministry)) is typed in by the annotator.

When an annotation is marked *onlyfactive=no*, additional attributes are used to characterize the private state. The *overall-strength* attribute is used to indicate the strength of the private state being expressed (considering the explicit private state or speech event phrase as well as everything inside its scope). It's value may range from low to extreme. The *on-strength*[4] attribute is used to measure the contribution made specifically by the explicit private state or speech event phrase. For example, the *on-strength* of "said" is typically neutral, the *on-strength* of "criticize" is typically medium, and the *on-strength* of "vehemently denied" is typically high or extreme. (As for all aspects of this annotation scheme, the annotators are asked to make these judgments in context.) A speech event that is *onlyfactive=yes* has *on-strength=neutral* and no *overall-strength*. Thus, there is no need to include the *overall-strength* and *on-strength* attributes for *onlyfactive=yes* annotations.

4.1.1 Implicit Speech Event Annotations

Implicit speech events posed a problem when we developed the annotation scheme. *Implicit speech events* are speech events in the discourse for which there is no explicit speech event phrase, and thus no obvious place to attach the annotation. For example, most of the writer's sentences do not include a phrase such as "I say." Also, direct quotes are not always accompanied by discourse parentheticals (such as ", she said"). Our solution was to add the *is-implicit* attribute to the annotation type for private states and speech events, which may then be used to mark implicit speech event annotations.

4.1.2 Minor Private States and Speech Events

Depending on its goals, an application may need to identify all private state and speech event expressions in a document, or it may want to find only those opinions and other private states that are significant and real in the discourse. By "significant", we mean that a significant portion of the contents of the private state or speech event are given within the sentence where the annotation is marked. By "real", we mean that the private state or speech event is presented as an existing event within the domain of discourse, e.g., it is not hypothetical. We use the term *minor* for private states and speech events that are not significant or not real. Annotators mark minor private state and speech event annotations by including the *minor* attribute.

The following sentences all contain one or more minor private states or speech events (highlighted in bold).

> **(11) Such wishful thinking** risks making the US an accomplice in the destruction of human rights. (not significant)

> **(12)** If the Europeans **wish** to influence Israel in the political arena... (in a conditional, so not real)

> **(13)** "And we are seeking **a declaration** that the British government **demands** that Abbasi should not face trial in a military tribunal with the death penalty." (not real, i.e., the declaration of the demand is just being sought)

> **(14)** The official **did not say** how many prisoners were on the flight. (not real because the saying event did not occur)

> **(15)** No one who has ever studied realist political science **will find this surprising**. (not real since a specific "surprise" state is not referred to; note that the subject noun phrase is attributive rather than referential (Donnellan, 1966))

4.2 Expressive Subjective Element Annotations

As with private state/speech event annotations, the *nested-source* attribute is included on every expressive subjective element annotation. In addition to marking the source of an expression, the *nested-source* is also functioning as a link. Within a sentence, the *nested-source* chains together all the pieces that together indicate the overall private state of a particular source.

[4] *on* is shorthand for a private state or speech event phrase

In addition to *nested-source*, the *strength* attribute is used to characterize expressive subjective element annotations. The *strength* of an expressive subjective element may range from low to extreme (see Table 1).

4.3 Exploratory Attributes

We are exploring additional attributes that allow an annotator to further characterize the type of attitude being expressed by a private state. An annotator may use the *attitude-type* attribute to mark an *onlyfactive=no* private state/speech event annotation or an expressive subjective element annotation as positive or negative. An *attitude-toward* attribute may also be included on private state/speech event annotations to indicate the particular target of an evaluation, emotion, etc.

5 Annotation Study

The data in our study consists of English-language versions of foreign news documents from FBIS, the U.S. Foreign Broadcast Information Service. The data is from a variety of publications and countries. To date, 252 articles have been annotated with the scheme described in section 4.

To measure agreement on various aspects of the annotation scheme, three annotators (A, M, and S) independently annotated 13 documents with a total of 210 sentences. None of the annotators are authors of this paper. The articles are from a variety of topics and were selected so that 1/3 of the sentences are from news articles reporting on objective topics (objective articles), 1/3 of the sentences are from news articles reporting on opinionated topics ("hot-topic" articles), and 1/3 of the sentences are from editorials.

In the instructions to the annotators, we asked them to rate the annotation difficulty of each article on a scale from 1 to 3, with 1 being the easiest and 3 being the most difficult. The annotators were not told which articles were objective or which articles were editorials, only that they were being given a variety of different articles to annotate.

We hypothesized that the editorials would be the hardest to annotate and that the objective articles would be the easiest. The ratings that the annotators assigned to the articles support this hypothesis. The annotators rated an average of 44% of the articles in the study as easy (rating 1) and 26% as difficult (rating 3). But, they rated an average of 73% of the objective articles as easy, and 89% of the editorials as difficult.

It makes intuitive sense that "hot-topic" articles would be more difficult to annotate than objective articles and that editorials would be more difficult still. Editorials and "hot-topic" articles contain many more expressions of private states, requiring an annotator to make more judgments than they would for objective articles.

5.1 Agreement for Expressive Subjective Element Annotations

For annotations that involve marking spans of text, such as expressive subjective element annotations, there are two issues that arise when measuring agreement between annotators. First, it is not unusual for two annotators to identify the same expression in the text, but to differ in how they mark the boundaries.[5] For example, both annotators A and M saw expressive subjectivity in the phrase, "such a disadvantageous situation." But, while A marked the entire phrase as a single expressive subjective element, M marked the individual words, "such" and "disadvantageous." For this work, we count overlapping annotations as matches.

The second issue is that annotators will identify different (but hopefully strongly overlapping) sets of expressions. Because of this, we need an agreement metric that can measure agreement between sets of objects. For expressive subjective element annotations (and later for private state/speech event annotations), we use the agr metric to measure agreement.

Let A and B be the sets of spans annotated by annotators a and b. agr is a directional measure of agreement that measures what proportion of A marked by a were also marked by b. Specifically, we compute the agreement of b to a as:

$$agr(a||b) = \frac{|A \text{ matching } B|}{|A|}$$

This measure of agreement corresponds to the notion of precision and recall as used to evaluate, for

[5]In the coding instructions, we did not attempt to define rules to try to enforce boundary agreement.

a	b	$agr(a\|\|b)$	$agr(b\|\|a)$	average
A	M	0.76	0.72	
A	S	0.68	0.81	
M	S	0.59	0.74	
				0.72

Table 2: Inter-annotator Agreement: Expressive subjective elements

a	b	$agr(a\|\|b)$	$agr(b\|\|a)$	average
A	M	0.75	0.91	
A	S	0.80	0.85	
M	S	0.86	0.75	
				0.82

Table 4: Inter-annotator Agreement: Explicitly-mentioned private states and speech events

example, named entity recognition. The $agr(a||b)$ metric corresponds to the recall if a is the gold-standard and b the system, and to precision, if they are reversed.

In the 210 sentences in the annotation study, the annotators A, M, and S respectively marked 311, 352 and 249 expressive subjective elements. Table 2 shows the pairwise agreement for these sets of annotations. For example, M agrees with 76% of the expressive subjective elements marked by A, and A agrees with 72% of the expressive subjective elements marked by M. The average agreement in Table 2 is the arithmetic mean of all six $agrs$.

We hypothesized that the stronger the expression of subjectivity, the more likely the annotators are to agree. To test this hypothesis, we measure agreement for the expressive subjective elements rated with a strength of medium or higher by at least one annotator. This excludes on average 29% of the expressive subjective elements. The average pairwise agreement rises to 0.80. When measuring agreement for the expressive subjective elements rated high or extreme, this excludes an average 65% of expressive subjective elements, and the average pairwise agreement increases to 0.88. Thus, annotators are more likely to agree when the expression of subjectivity is strong. Table 3 gives examples of expressive subjective elements that at least one annotator rated as extreme.

5.2 Agreement for Private State/Speech Event Annotations

For private state and speech event annotations, we again use agr to measure agreement between the sets of expressions identified by each annotator. The three annotators, A, M, and S, respectively marked 338, 285, and 315 explicit expressions of private states and speech events. Implicit speech events for the writer of course are excluded. Table 4 shows the pairwise agreement for these sets of annotations.

The average pairwise agreement for explicit private state and speech event expressions is 0.82, which indicates that they are easier to annotate than expressive subjective elements.

5.3 Agreement for Attributes

In this section, we focus on the annotators' agreement for judgments that reflect whether or not an opinion, emotion, or other private state is being expressed. We consider these judgments to be at the core of the annotation scheme. Two attributes, *onlyfactive* and *on-strength*, carry information about whether a private state is being expressed.

For *onlyfactive* judgments, we measure pairwise agreement between annotators for the set of private state and speech event annotations that both annotators identified. Because we are now measuring agreement over the same set of objects for each annotator, we use Kappa (κ) to capture how well the annotators agree.

Table 5 shows the contingency table for the *onlyfactive* judgments made by annotators A and M. The Kappa scores for all annotator pairs are given in Table 7. For their *onlyfactive* judgments, i.e., whether or not an opinion or other private state is being expressed, the annotators have an average pairwise Kappa of 0.81. Under Krippendorf's scale (Krippendorf, 1980), this allows for definite conclusions.

With many judgments that characterize natural language, one would expect that there are clear cases as well as more difficult to judge borderline cases. The agreement study indicates that this is certainly true for private states. In terms of our annotations, we define an explicit private state or speech event to be *borderline-onlyfactive* if 1) at least one annotator marked the expression *onlyfactive=no*, and 2) neither annotator character-

mother of terrorism
if the world has to rid itself from this menace, the perpetrators across the border had to be dealt with firmly
indulging in blood-shed and their lunaticism
ultimately the demon they have reared will eat up their own vitals

Table 3: Extreme strength expressive subjective elements

		Tagger M: Yes	Tagger M: No
Tagger A	Yes	$n_{yy} = 181$	$n_{yn} = 25$
	No	$n_{ny} = 12$	$n_{nn} = 252$

Table 5: A & M: Agreement for *onlyfactive* judgments

		Tagger M: Yes	Tagger M: No
Tagger A	Yes	$n_{yy} = 181$	$n_{yn} = 8$
	No	$n_{ny} = 11$	$n_{nn} = 224$

Table 6: A & M: Agreement for *onlyfactive* judgments, *borderline-onlyfactive* cases removed

ized its *overall-strength* as being greater than low. In Table 6 we give the contingency table for the *onlyfactive* judgments made by annotators A and M, excluding *borderline-onlyfactive* expressions. Note that removing such expressions removes agreements as well as disagreements. *Borderline-onlyfactive* expressions on average comprise only 10% of the private state/speech event annotations. When they are removed, the average pairwise Kappa climbs to 0.89.

In addition to the *onlyfactive* judgment, using *on-strength* we can measure if the annotators agree as to whether an explicit private state or speech event phrase by itself expresses a private state. Specifically, we measure if the annotators agree that an expression is neutral, i.e., does not indicate a private state. Recall that *onlyfactive=yes* annotations are *on-strength=neutral*. Implicit annotations are excluded when measuring *on-strength* agreement.

	All Expressions		Borderline Removed		
	κ	agree	κ	agree	% removed
A & M	0.84	0.91	0.94	0.96	10
A & S	0.84	0.92	0.90	0.95	8
M & S	0.74	0.87	0.84	0.92	12

Table 7: Pairwise Kappa scores and overall percent agreement for *onlyfactive* judgments

	All Expressions		Borderline Removed		
	κ	agree	κ	agree	% removed
A & M	0.81	0.91	0.93	0.97	22
A & S	0.74	0.87	0.92	0.96	17
M & S	0.67	0.83	0.90	0.95	18

Table 8: Pairwise Kappa scores and overall percent agreement for *on-strength* neutral judgments

The pairwise agreement results for the annotators' *on-strength* neutral judgments are given in Table 8. For *on-strength* neutral judgments, annotators have an average pairwise Kappa of 0.74. As with the *onlyfactive* judgments, there are clearly borderline cases. We define an expression to be *borderline-low* if 1) at least one annotator marked the expression *onlyfactive=no*, and 2) neither annotator characterized its *on-strength* as being greater than low. When *borderline-low* expressions are removed, the pairwise Kappa increases to 0.92.

5.4 Agreement for Sentences

To compare our results to those of earlier work that evaluated the agreement of sentence-level subjectivity annotations (Wiebe et al., 1999), we define sentence-level classifications in terms of our lower-level annotations as follows. First, we exclude explicit private state/speech event expressions that the annotators agree are minor. Then, if an annotator marked one or more *onlyfactive=no* expressions in the sentence, we consider the annotator to have judged the sentence to be subjective. Otherwise, we consider the annotator to have judged the sentence to be objective.

The pairwise agreement results for these derived sentence-level annotations are given in Table 9. The average pairwise Kappa for sentence-level agreement is 0.77, 8 points higher than the sentence-level agreement reported in (Wiebe et al., 1999). Our new results suggest that adding detail to the annotation task can can help annotators perform more reliably. Note that the agree-

	All Sentences		Borderline Removed		
	κ	agree	κ	agree	% removed
A & M	0.75	0.89	0.87	0.95	11
A & S	0.84	0.94	0.92	0.97	8
M & S	0.72	0.88	0.83	0.93	13

Table 9: Pairwise Kappa scores and overall percent agreement for derived sentence-level judgments

ment is lower than that for *onlyfactive* judgments (Table 7) because explicit private-state and speech event expressions upon which the annotators did not agree are now included.

As with the *onlyfactive* and *on-strength* neutral judgments, we again test agreement when borderline cases are removed. We define a sentence to be *borderline* if 1) at least one annotator marked at least one expression *onlyfactive=no*, and 2) neither annotator marked an *overall-strength* attribute as being greater than low. When *borderline* sentences are removed, the average Kappa increases to 0.87.

6 Conclusions

In this paper, we presented a detailed scheme for the annotation of opinions and other private states in the news and other discourse. For the aspects of this annotation scheme that indicate whether a private state is expressed, our three annotators have strong pairwise agreement, as measured by Cohen's Kappa.

One interesting area explored in this paper is the effect of borderline cases on inter-annotator agreement. We created a number of objective definitions of borderline cases, based on the strengths indicated by the annotators, and found that removing these borderline cases always results in high agreement values. This shows that the annotators agree strongly about which are the clear cases of subjectivity.

We have also shown that lower-level subjectivity annotations, such as those presented in this paper, may be used to produce higher-level subjective sentence annotations. In current research, we are using these higher-level annotations to evaluate subjective sentence classifiers, which we hope will be useful for enhancing natural language processing applications such as information extraction, summarization, and question answering systems.

There are characteristics of private state expressions not yet included in our scheme that would be useful for NLP applications. We believe the scheme is extendable, and hope that other groups will build on it.

References

A. Banfield. 1982. *Unspeakable Sentences*. Routledge and Kegan Paul, Boston.

C. Cardie, J. Wiebe, T. Wilson, and D. Litman. 2003. Combining low-level and summary representations of opinions for multi-perspective question answering. In *Working Notes - New Directions in Question Answering (AAAI Spring Symposium Series)*.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Meas.*, 20:37–46.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. 2002. GATE: A framework and graphical development environment for robust nlp tools and applications. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Keith Donnellan. 1966. Reference and definite descriptions. *Philosophical Review*, 60:281–304.

M. Fludernik. 1993. *The Fictions of Language and the Languages of Fiction*. Routledge, London.

K. Krippendorf. 1980. *Content Analysis: An Introduction to its Methodology*. Sage Publications, Beverly Hills.

M. Marcus, Santorini, B., and M. Marcinkiewicz. 1993. Building a large annotated corpus of English: The penn treebank. *Computational Linguistics*, 19(2):313–330.

R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik. 1985. *A Comprehensive Grammar of the English Language.* Longman, New York.

William Rapaport. 1986. Logical foundations for belief representation. *Cognitive Science*, 10:371–422.

D. Stein and S. Wright, editors. 1995. *Subjectivity and Subjectivisation*. Cambridge University Press, Cambridge.

T.A. van Dijk. 1988. *News as Discourse*. Lawrence Erlbaum, Hillsdale, NJ.

J. Wiebe, R. Bruce, and T. O'Hara. 1999. Development and use of a gold standard data set for subjectivity classifications. In *Proc. 37th Annual Meeting of the Assoc. for Computational Linguistics (ACL-99)*, pages 246–253, University of Maryland, June. ACL.

J. Wiebe, E. Breck, C. Buckley, C. Cardie, P. Davis, B. Fraser, D. Litman, D. Pierce, E. Riloff, T. Wilson, D. Day, and M. Maybury. 2003. Recognizing and organizing opinions expressed in the world press. In *Working Notes - New Directions in Question Answering (AAAI Spring Symposium Series)*.

J. Wiebe. 1994. Tracking point of view in narrative. *Computational Linguistics*, 20(2):233–287.

A Sample Annotations

The following is the first sentence from an article about the 2002 presidential election in Zimbabwe. The article appeared on March 15, 2002 in the newspaper, *Dawn*.

> Western countries were left frustrated and impotent after Robert Mugabe formally declared that he had overwhelmingly won Zimbabwe's presidential election.

There are three private state/speech event annotations and one expressive subjective element annotation in this sentence. The annotations, including their attributes, are listed below:

Speech Event: implicit
nested-source = (writer)
onlyfactive = yes

Private State: were left frustrated
nested-source = (writer, Western countries)
onlyfactive = no
overall-strength = medium
on-strength = medium

Speech Event: formally declared:
nested-source = (writer, Mugabe)
onlyfactive = no
overall-strength = medium
on-strength = neutral

Expressive Subjective Element: overwhelmingly:
nested-source = (writer, Mugabe)
strength = medium

Figure 1 shows how these annotations appear inside the GATE annotation tool. Additional annotated examples can be found with the on-line GATE annotation instructions, http://www.cs.pitt.edu/mpqa/opinion-annotations/gate-instructions.

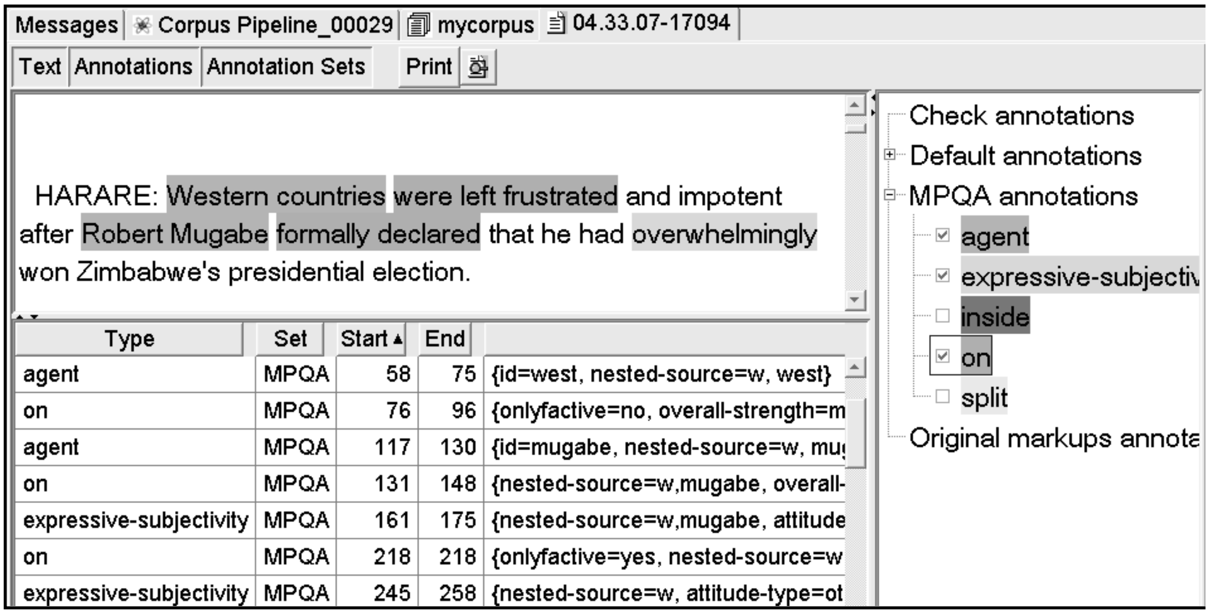

Figure 1: Example of annotations in GATE